nature genetics

PERSPECTIVE

https://doi.org/10.1038/s41588-021-00866-3



# Advancing crop genomics from lab to field

Michael D. Purugganan[1,2] ✉ and Scott A. Jackson[3]

**Crop genomics remains a key element in ensuring scientific progress to secure global food security. It has been two decades since the sequence of the first plant genome, that of *Arabidopsis thaliana*, was released, and soon after that the draft sequencing of the rice genome was completed. Since then, the genomes of more than 100 crops have been sequenced, plant genome research has expanded across multiple fronts and the next few years promise to bring further advances spurred by the advent of new technologies and approaches. We are likely to see continued innovations in crop genome sequencing, genetic mapping and the acquisition of multiple levels of biological data. There will be exciting opportunities to integrate genome-scale information across multiple scales of biological organization, leading to advances in our mechanistic understanding of crop biological processes, which will, in turn, provide greater impetus for translation of laboratory results to the field.**

The world continues to be challenged by food insecurity[1,2]. After decades of steady decline, the prevalence of undernourishment in the world reversed course starting in 2015 and has once again begun to rise. Undernourishment currently hovers at just under 9% worldwide, but is projected to grow to 9.8% by 2030, when nearly 850 million people are predicted to experience hunger[1]. Moreover, agriculture continues to have an enormous resource footprint, taking up 38% of the Earth's land surface, consuming ~70% of the world's freshwater and using 1.2% of global energy.

The root causes of food insecurity amid outsized agricultural resource consumption are numerous and include human population growth (which should stabilize at 9–11 billion by the end of the century), climate change, urbanization, deterioration of agricultural land, reliance on carbon-economy-based chemical inputs and increasing water scarcity[1,2]. Feeding an increasingly hungry world remains one of the key challenges facing humanity, and the United Nations has set a Zero Hunger Target by 2030. It is imperative to meet this task in a sustainable manner, by ensuring growth in crop yields in the face of deteriorating environments while reducing the resources necessary to feed a burgeoning world population.

Success in this global endeavor will require a systems-based approach that incorporates new data-driven farming methods, novel sustainable practices and improved crop cultivars; in this, genomics provides foundational tools and biological insights for 21st century agriculture. It was more than 20 years ago that the first whole genome sequence of a plant—that of *Arabidopsis thaliana*—was released[3], and in 2002 rice was the first crop to have its genome sequenced (Fig. 1)[4,5]. Over the last two decades since these first plant genome sequences were released, genomic science has helped enhance plant breeding efforts, allowing increasing yields, providing resilience to environmental and pathogen stresses, and developing novel varieties[6,7]. As we move into this century, genomics will continue to play a pivotal role in reshaping crop biology to meet current and future needs[6], just as genetics helped drive crop improvement over the last century.

Advances in technologies and approaches have extended the field even further, allowing the sequencing of complex crop genomes, including the large hexaploid 17-gigabase (Gb) wheat genome[8]. Genome sequencing, including long-read single molecule sequencing, has now become routine[9,10]. New tools have allowed genome-level analyses of epigenomic information[11,12], including the three-dimensional conformation of the genome in the nucleus[13], and large-scale transcriptome[14], metabolome[15] and proteome[16] information are now readily obtained. Robotic techniques underlie high-throughput phenotyping studies[17,18], even in real-world field environments[19], sometimes coupled with drone[20] or even satellite imagery[21]. These techniques also benefit from new computational methods to analyze large and disparate datasets, together with contemporary algorithms that can incorporate machine learning[22–24] and artificial intelligence[25]. Meanwhile, technologies such as CRISPR–CAS9 allow greater precision in editing genomes for evaluating gene function, for engineering new genomes for crop improvement[26,27] and even for de novo domestication of novel crop species[28].

These new technologies and approaches are setting the stage for novel lines of inquiry and exciting horizons for crop genomics in the coming years. A crucial element will be the integration of multiple streams of disparate data to develop new insights into crop biology, with downstream applications to agriculture. Moreover, many of these technologies have been developed in other fields, particularly in biomedical genomics, and will likely impact crop genomics research, as they have in the past. And while most of this technological progress has been reviewed elsewhere, synthesizing these perspectives points to new directions in crop genomics, helping us better understand crop plant biology and in so doing not only advance biological understanding but also address global issues in agriculture.

## A new era for genome sequencing

Sequencing has always been the cornerstone of genomic science, and in the last two decades the genomes of about 10 cereal crop and >100 vegetable and fruit species have been sequenced (Fig. 1)[29]. Today, it has never been easier to decode and assemble crop genomes; innovations in short-read sequencing technologies (such as improved linked-read sequencing)[30], coupled with the increase in ease and decrease in costs of long-read sequencing (including single molecule real time sequencing[31] and nanopore methods[32,33]), have improved genome assemblies, making it possible to routinely tackle whole genome sequencing projects[34].

These advances have democratized the ability to develop new reference genome sequences either in draft form or in high-quality near-complete assemblies[35]. The goal of a reference genome is

[1]Center for Genomics and Systems Biology, New York University, New York, NY, USA. [2]Center for Genomics and Systems Biology, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. [3]Bayer Crop Science, Chesterfield, MO, USA. ✉e-mail: mp13@nyu.edu

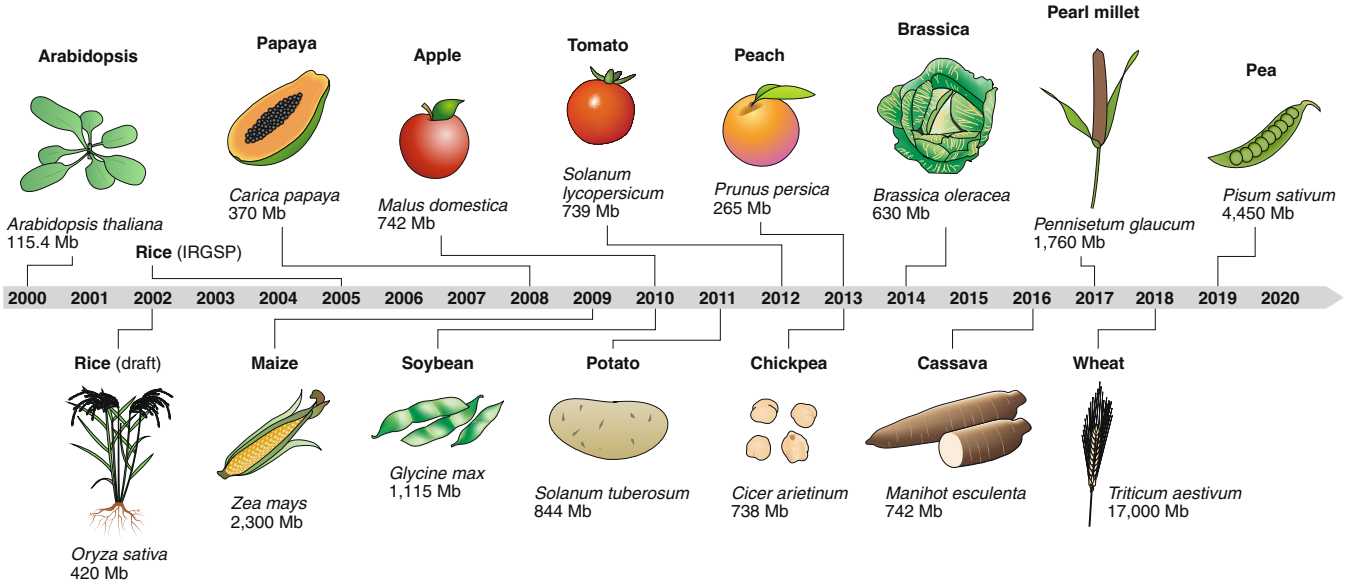


**Fig. 1 | Timeline of release of genome sequences for key crop species.** The year of the release of the genome sequence, as well as the either estimated or assembled size of the genome, is indicated. *A. thaliana* was included given its landmark status as the first plant genome sequenced. The crop species depicted were chosen either for their agricultural importance or to depict a wide variety of food species. For rice, the first crop genome sequenced, we indicate the release of the first drafts as well as the map-based sequences of the genome by the International Rice Genome Sequencing Project (IRGSP). Mb, megabases.

telomere-to-telomere contiguity to fully characterize a genome sequence; this completeness is made possible by technologies such as long-read sequencing, genome-wide contact and restriction maps[36], high-throughput chromosome conformation capture[37] and optical mapping[38,39]. These approaches are coupled with improved computational tools for genome assembly (reviewed in ref. [10]), and have led to higher-quality reference genome sequences across a broad range of species. Moreover, new technologies together with different sequencing strategies—including genome complexity reduction[40], diploid progenitor sequencing[41] and individual chromosome sequencing[42]—have made accessible even species with large genomes, such as maize (2.3 Gb)[43,44], barley (5.3 Gb)[45] and pea (4.45 Gb)[46]. Indeed, even polyploid crop genomes such as allotetraploid cotton (2.5 Gb)[47], hexaploid wheat (~17 Gb)[8] and octoploid sugarcane (3.13 Gb)[48] can now be explored with greater facility, which will expand the scope of genome studies for key crop species and enable further comparative and genetic mapping studies.

Although there are approximately 20,000 plant species that are edible to humans[49], global food production focuses on a few large-scale crops (for example, rice, wheat, maize and soybean), and the genomics of many domesticated food species remain unexplored. Given the reduction in costs and ease of sequencing and assembly, genome resources can now be made available for niche, local species—so-called orphan crops[50]—that may not have global importance but are nevertheless crucial for local economies and food systems. Moreover, focused attention on landraces/traditional varieties and wild relatives of the major crops can now be investigated in greater detail, expanding the gene pool available to breeders for the improvement of domesticated species[51,52]. These less well-known orphan crops, and genome sequencing of crop landraces and wild relatives, may prove particularly critical in crop breeding for future environments, as several of these populations and species are adapted to unique and even stressful conditions, and may be increasingly important for climate change adaptation.

The untapped potential of genetic information in crop landraces deserves particular attention. Given the ease in current sequencing approaches, the time has also come to systematically unlock the genetic diversity found in large collections of crop landraces and cultivars kept at seed banks throughout the world[53,54]. It is estimated that ~7.4 million seed accessions can be found in about 1,700 germplasm collections worldwide[55], which are unparalleled resources both for advancing genetic knowledge as well as for breeding efforts. The 3,000 Rice Genome Project, which a few years ago released genome sequence information for 3,010 rice varieties[56], has already proved invaluable; there are now calls to sequence all of the ~128,000 varieties kept at the seed bank of the International Rice Research Institute, to develop a 'digital genebank' and help identify rare alleles of key agronomic genes[57]. There is also a 3,000 Chickpea Genome Sequencing Initiative to advance genetic mapping efforts in this key developing country crop species[58], and it is likely that more large-scale endeavors such as these will proliferate[59,60]. An area of future exploration using these large-scale projects will be to identify and exploit alleles favored by evolution in response to past climate change or adaptation to local environments. This approach will by necessity integrate population, evolutionary and functional genomics with past climate modeling and landscape genomics; such work has already begun in *A. thaliana*[61–63] as well as in rice[64], maize[65], sorghum[66], pearl millet[67] and fonio[68]; nevertheless, there is much to do to fully realize the promise of this approach in crop improvement.

Finally, the ability to develop multiple high-quality reference genome sequences for each species will help untangle the paradox of the pan-genome[69,70]. It is now clear that a species genome consists of both a core genome, whose sequence is found across the entire species, and an accessory genome comprising dispensable sequences found in some but not all individuals of the species. In rice, for example, 38% of genes are thought to be dispensable[71], which is similar to the fraction of genes (~33%) that are thought to have variable presence across maize[72]. Pan-genomes of other crop species have also been examined, including hexaploid wheat[73], barley[74], soybean[75] and tomato[76]. The biological importance of the pan-genome remains unclear—how do crop plants maintain function despite variable gene content, and to what extent are presence and/or absence of genes associated with crop adaptive variation? There certainly is good evidence to demonstrate the importance of

pan-genomic variation in key crop genes and phenotypes. In soybean, for example, pan-genome variation appears to be associated with genes for seed luster, seed pigmentation and flowering time[75], while in wheat there is variation in gene content among cultivars in disease and insect resistance genes[73]. In tomato, a large structural variant at the promoter region of *TomLoxC* is implicated in fruit flavor differences[76]. These initial studies demonstrate the key role that presence/absence variation and structural polymorphisms found in crop pan-genomes may play in adaptation and diversity, and are an exciting area of research both to illuminate plant biology and to improve crops.

## Expanding genetic mapping

The last two decades have seen the identification of key genes underlying agricultural traits, using quantitative trait locus (QTL) mapping and genome-wide association studies (GWASs) combined with molecular genetic analyses that enable map-based cloning and functional investigation of important loci. Over the last 10 years, there have been >1,000 published GWASs in more than 20 crop species[77], and these numbers will clearly increase. This area of crop genomics research has matured considerably, utilizing ever larger populations and nested association mapping panels, new computational approaches and ancillary data (for example, epigenomic marks) that help deliver higher genetic resolution and provide functional characterization. Indeed, GWASs in crop species have become almost routine, and have added considerably to our understanding of the genetic architecture of crop traits.

Our understanding of the genetic architecture of crop traits, however, remains incomplete. QTL mapping and GWAS methods have helped identify genes of moderate-to-large effect, but it remains challenging to genetically dissect strongly polygenic traits. Moreover, the focus has been on additive genes, while identifying loci associated with epistatic and gene-by-environment interactions[78] has proven much more difficult. New methods will be necessary if we are to make progress towards a more complete molecular specification of trait genetic architectures in crops in all their complexities.

Current GWAS mapping methods can also often overlook rare alleles that are found in only one or a few crop individuals in a population. Indeed, causal alleles for key traits in rice identified by QTL mapping—such as grain size (*GS3* (ref. [79])) and *qGL3* (ref. [80])) and flowering time (*GHd7*)[81]—are found at <2% frequency in the population. Rare alleles may also result from the movement of transposable elements, abundant in plant genomes, that correlate with phenotypic changes such as grain width in rice[82] but may not be captured using traditional mapping approaches. Greater power in GWAS mapping can be provided by even larger mapping populations, and the move to sequence greater numbers of cultivars from seed banks (see above) may help identify these rare but agronomically valuable alleles.

Attention should also focus on mapping genes in crop wild relatives. These wild species fend off pathogen and pest attacks, poor soils, water deficits and weather extremes in their natural wild habitats, and can potentially offer new genes for crop improvement[51,52]. There is as yet relatively little genetic and GWAS mapping in crop wild relatives, and this could potentially be a productive area of research in the coming years.

Finally, we expect advances in genetic mapping methodologies. There already are new algorithms, such as FarmCPU[83], that allow one to conduct GWAS mapping of loci that may be important in local adaptation but are inevitably confounded with population structure. Other methods, such as BLINK[84], can provide greater power to identify genes by using very dense SNP marker data, thus increasing the resolution of GWAS peaks. Methods of GWAS mapping even in the absence of a reference genome sequence[85] may allow more crop species to be analyzed, particularly orphan crops which suffer from a dearth of genomic resources. Finally, specialized mapping populations, of which multiparent advanced generation intercross lines[86,87], nested mapping populations[88] and even mutagenized populations[89] have proved useful in increasing the power of genetic mapping, and other innovative mapping strategies may be forthcoming.

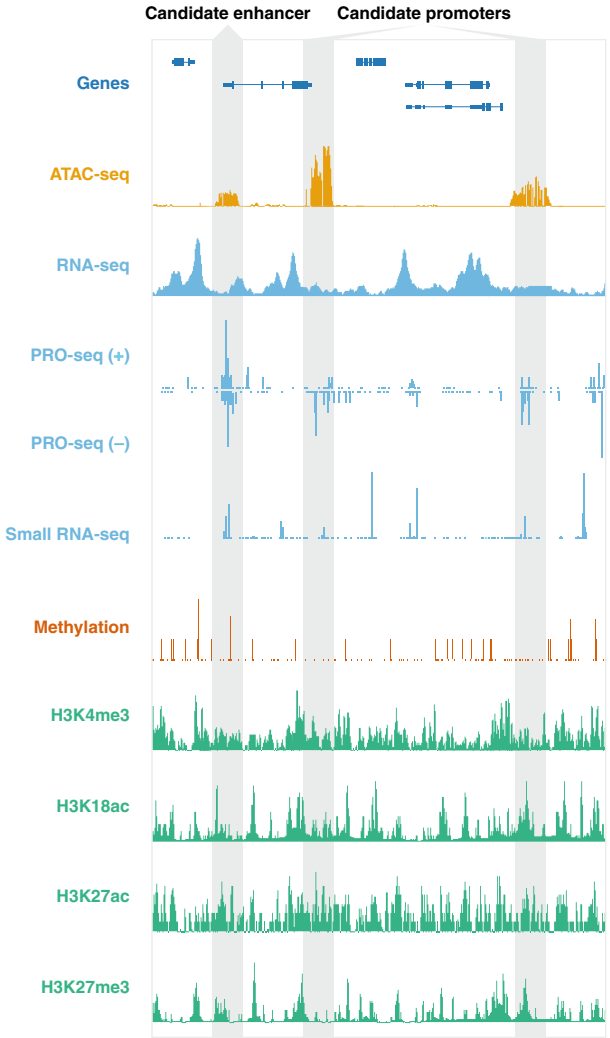


**Fig. 2 | Maps of genome-wide functional genomic and epigenomic information.** Going beyond sequence data, the ability to assay for various functional and biochemical marks provide new tools for understanding gene function and regulation. An example is shown from rice[97], with the gene model shown on top and levels of various features shown below, including RNA-sequencing (RNA-seq) levels, open chromatin as assayed by assay for transposase-accessible chromatin using sequencing (ATAC-seq), methylation levels, and various histone methylation and acetylation modifications. One can visualize the open chromatin (ATAC-seq track) and histone modification (H3K4me3 track) associated with transcribed genes (RNA-seq track), as well as bi-directional nascent transcription (precision run-on sequencing (PRO-seq)[125] track) that signals possible enhancer sequences.

## Systems genomics of crops

While there have been successes in identifying critical genes for agronomic improvement, the mechanistic understanding of gene functions and how they specify agricultural phenotypes continues to lag. In the coming years, more attention should be paid to unraveling molecular mechanisms that underlie traits and forge the links

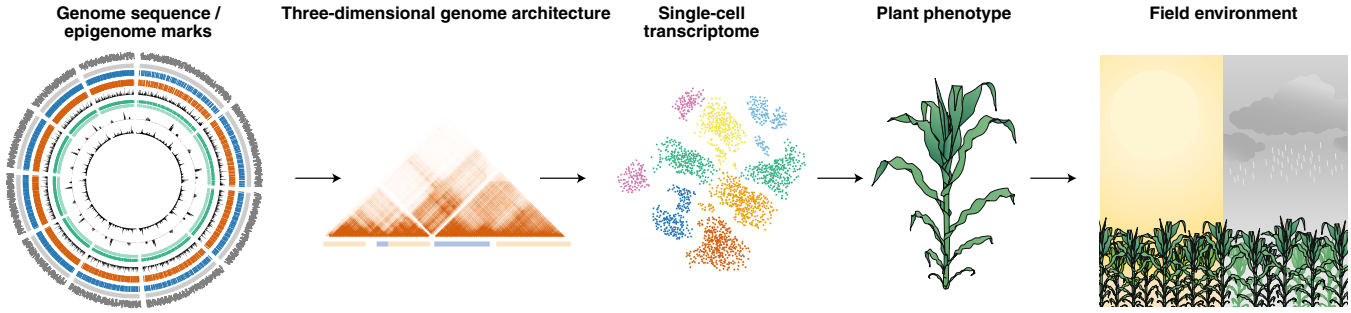


**Fig. 3 | Multiple levels of functional genomic, epigenomic organismal and ecosystem information.** There is a need to integrate information across a hierarchy of biological scales. As an example, from left to right, we can have genome sequence and epigenomic marks, three-dimensional chromosome conformation, gene expression data (here hypothetically clustered at the single-cell transcriptome level), tissue/organ and organismal phenotypes, and field environmental/ecosystem factors. The challenge is how to practically and conceptually connect across these different scales.

between genotype and phenotype, to employ this knowledge in furthering crop improvement.

Dissecting genetic mechanisms underlying phenotypes will have to embrace complexity, and in this light systems genomics approaches will continue to expand[90,91]. A systems approach examines multiple levels of organization, from genome sequences to single-cell transcriptomics[92,93] to developmental trajectories to multispecies interactions (including pathogen,[94] insect[95] and microbiome[96]) and fluctuating multi-environmental perturbations. This will entail collecting various functional genomic and epigenomic information on a genome-wide scale, which can be used to infer gene function and regulation (Fig. 2)[97]. The challenge is to determine how best to obtain meaningful data across different organizational and temporal scales and integrate these to gain a mechanistic appreciation to how they specify organismal traits.

One increasingly exciting area has been the ability to assay functional genomic and epigenomic data at the single-cell level[98]. The ability to examine how transcriptomes of discrete cells evolve over developmental time[93,99] or under stress[100], for example, can provide greater resolution in studies of plant cell differentiation and physiological response. Moreover, there is also increasing attention to various epigenomic features of the plant genome, including the high-dimensional reconstruction of chromosomal contacts[13,101]. Together, these approaches are providing new insights into the nature of cell identity and gene regulation that can eventually inform high-precision efforts for genome editing, both to learn about underlying mechanisms and for crop improvement.

Understanding the behavior of crops in the field remains a vital and largely unexplored area[19,102,103]. How crop plants respond to environmental fluctuations that occur at different time scales (from sunflecks that fluctuate in seconds to seasonal patterns that vary monthly) needs to be studied to obtain a clearer picture on how plants maintain function amid variable environmental signals. There are opportunities to try to integrate field environmental, climatic, historical/evolutionary and phenotypic information in reconstructing crop adaptation and evolution in specific environmental niches. This may benefit from industry–academic collaboration, as industrial researchers have decades-long, phenotypically rich datasets from field trials around the world that academic researchers could exploit to examine mechanisms underlying crop adaptation.

Finally, as more systems genomics approaches are pursued, the amount, quality and type of genomic data that are now routinely acquired in crop species will continue to grow. In principle, we now (or will soon have) data to examine multiple levels of biological relevance (Fig. 3)[104,105], and the current deluge of data calls on us to address how to mine these enormous datasets for new biology, and how to integrate data across multiple scales to gain new biological insights. In particular, computational methods that can examine diverse information sets need to be developed. Computational analysis and mathematical modeling across organizational scales are inherently difficult. Newer approaches in data science and artificial intelligence such as machine learning can help uncover patterns in large-scale and disparate data types[22–25], but do not immediately reveal biological mechanism. Model-based analyses could be more insightful, but will require previous knowledge to help construct appropriate models[104]. Nevertheless, these challenges are also opportunities for discovery, and in crop genomics could pay future dividends in modern agricultural advances.

## From genome to the field—advancing translation

While genomics has accelerated the pace of genetic discovery, there remains a gap in the application of these findings to breeding programs. It has now become straightforward to discover GWAS associations, for example, but deploying these discoveries in the development of new crop varieties has not been as robust. There are, of course, great examples of moving the results of genetic mapping to the field, including submergence tolerance in rice[106]. Several questions, however, continue to stymie translational progress. For instance, how do you place a breeding value on GWAS peaks[107]? How do you routinely incorporate GWAS results into genomic selection/prediction models[108–110]? Tackling these questions will also require attention to model and algorithm development, innovatively integrating various types of data (genome, phenotype, functional, epigenomic and so on) and taking into account nonlinear interactions such as genotype-by-environment effects[9,111] in genomic selection/genomic prediction approaches to advance crop breeding goals.

Historically, there has been a mechanistically blind approach to crop improvement, with a focus on phenotypic observations and/or quantitative genetic breeding values. With some exceptions, there has been little functional understanding of how agronomically important traits or phenotypes develop at the molecular level. On the other hand, molecular geneticists have made advances in unraveling the genetic basis of certain traits (flowering time, root development, photosynthesis, stress response and so on), and the task is how to routinely incorporate this knowledge for crop improvement[112].

We should recognize that the evolutionary processes of domestication and crop diversification have led to both constraints and opportunities that can inform breeding efforts. For example, reduced effective population sizes and genetic hitchhiking from positive selection inevitably lead to increases in the levels of deleterious polymorphisms in crop populations while also limiting the variation available for further crop improvement. There has been

growing appreciation of the extent and role of deleterious segregating polymorphisms in holding back crop yields[113–116] and it has been suggested that breeding programs can aim to identify and purge deleterious mutations from domesticated populations[97,117]. Conversely, crop diversification across multiple environments during evolution results in local adaptation, which could provide important genetic material to help in developing crop varieties suitable for niche ecosystems[118].

Genome editing, using CRISPR–CAS9 (refs. [26,27]) as well as other engineered nucleases[119] such as zinc-finger nucleases[120] and transcription activator-like effector nucleases[121], is a technology that will help push crop breeding efforts and further provide linkages between mechanistic understanding of gene action and agricultural outcomes[122]. Some of these molecular tools have been used to increase resistance against bacterial blight in rice[123] or develop compact tomato plants for urban agriculture[124]. As these technologies are refined to engineer subtler mutational effects in crop genomes, it may be possible to deliver a wider range of phenotypes useful to agriculture.

## Conclusion

Crop genomics has been a key driver of agricultural advances across the first two decades of the 21st century. The ability to sequence genomes and assay multiple layers of functional genomic and epigenomic information has proven crucial in helping obtain a clearer understanding of plant biology. There are exciting opportunities to integrate multiple levels of data—for example, integrating gene expression, metabolome and environmental data; or three-dimensional chromosomal conformation with sequence data and evolutionary information. As genomic technologies and computational approaches continue to move forward, we should see an increasing ability to explore important plant traits that can translate to improvements in farmers' fields. Genomic science will continue to provide important insights and tools to help us minimize food insecurity and to lay the foundation for a sustainable agricultural system to feed the world.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Received: 27 October 2020; Accepted: 22 March 2021;
Published online: 6 May 2021

## References

1. Food and Agriculture Organization. *The State of Food Security and Nutrition in the World 2020* (FAO, 2019).
2. Foley, J. et al. Solutions for a cultivated planet. *Nature* **478**, 337–342 (2011).
3. The Arabidopsis Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
4. Goff, S. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
5. Yu, J. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
6. Bevan, M. et al. Genomic innovation for crop improvement. *Nature* **543**, 347–354 (2017).
7. Briggs, S. P. Plant genomics: more than food for thought. *Proc. Natl Acad. Sci. USA* **95**, 1986–1988 (1998).
8. International Wheat Genome Sequencing Consortium (IWGSC) et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018)
9. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
10. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
11. EPIC Planning Committee. Reading the second code: mapping epigenomes to understand plant growth, development, and adaptation to the environment. *Plant Cell* **24**, 2257–2261 (2012).
12. Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
13. Ricci, W. et al. Widespread long-range *cis*-regulatory elements in the maize genome. *Nat. Plants* **5**, 1237–1249 (2019).
14. Ramírez-González, R. H. et al. The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).
15. Chen, W. et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* **46**, 714–721 (2014).
16. Mergner, J. et al. Mass-spectrometry-based draft of the *Arabidopsis* proteome. *Nature* **579**, 409–414 (2020).
17. Furbank, R. & Tester, M. Phenomics—technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* **16**, 635–644 (2011).
18. Araus, L. et al. Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* **23**, 451–466 (2018).
19. Zaidem, M. L., Groen, S. C. & Purugganan, M. D. Evolutionary and ecological functional genomics, from lab to the wild. *Plant J.* **97**, 40–55 (2019).
20. Tattaris, M., Reynolds, M. P. & Chapman, S. C. A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding. *Front. Plant Sci.* **7**, 1131 (2016).
21. Clevers, J., Kooistra, L. & van den Brande, M. Using Sentinel-2 data for retrieving LAI and leaf and canopy chlorophyll content of a potato crop. *Remote Sens.* **9**, 405 (2017).
22. Ma, C., Zhang, H. H. & Wang, X. Machine learning for Big Data analytics in plants. *Trends Plant Sci.* **19**, 798–808 (2014).
23. Esposito, S. et al. Applications and trends of machine learning in genomics and phenomics for next-generation breeding. *Plants* **9**, 34 (2020).
24. Wang, H., Cimen, E., Singh, N. & Buckler, E. Deep learning for plant genomics and crop improvement. *Curr. Opin. Plant Biol.* **54**, 34–41 (2020).
25. Harfouche, A. et al. Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends Biotech.* **37**, 1217–1235 (2019).
26. Belhaj, K. et al. Editing plant genomes with CRISPR/Cas9. *Curr. Opin. Biotech.* **32**, 76–84 (2015).
27. Chen, K. et al. CRISPR/Cas genome editing and precision plant breeding in agriculture. *Ann. Rev. Plant Biol.* **70**, 667–697 (2019).
28. Fernie, A. R. & Yan, J. De novo domestication: an alternative route toward new crops for the future. *Mol. Plant* **12**, 615–631 (2019).
29. Chen, F. et al. Genome sequences of horticultural plants: past, present, and future. *Hort. Res.* **6**, 112 (2019).
30. Ott, A. et al. Linked read technology for assembling large complex and polyploid genomes. *BMC Genomics* **19**, 651 (2018).
31. Roberts, R., Carneiro, M. & Schatz, M. The advantages of SMRT sequencing. *Genome Biol.* **14**, 405 (2013).
32. Branton, D. et al. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
33. Belser, C. et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879–887 (2018).
34. Choi, J. Y. et al. Nanopore sequencing-based genome assembly and evolutionary genomics of *circum*-basmati rice. *Genome Biol.* **21**, 21 (2020).
35. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
36. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
37. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
38. Levy-Sakin, M. & Ebenstein, Y. Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Curr. Opin. Biotechnol.* **24**, 690–696 (2013).
39. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
40. Rabinowicz, P. D. et al. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* **23**, 305–308 (1999).
41. Bertioli, D. J. et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **48**, 438–446 (2016).
42. Kopecký, D. et al. Flow sorting and sequencing meadow fescue chromosome 4F. *Plant Physiol.* **163**, 1323–1337 (2013).
43. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
44. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
45. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
46. Kreplak, J. et al. A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* **51**, 1411–1422 (2019).

47. Wang, M. et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* **51**, 224–229 (2019).
48. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
49. Schaal, B. Plants and people: our shared history and future. *Plants People Planet* **1**, 14–19 (2019).
50. Varshney, R. et al. Can genomics boost productivity of orphan crops? *Nat. Biotech.* **30**, 1172–1176 (2012).
51. Brozynska, M., Furtado, A. & Henry, R. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotech. J.* **14**, 1070–1085 (2016).
52. Dempewolf, H. et al. Past and future use of wild relatives in crop breeding. *Crop Sci.* **57**, 1070–1082 (2017).
53. Mascher, M. et al. Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* **51**, 1076–1081 (2019).
54. McCouch, S. et al. Mobilizing crop biodiversity. *Mol. Plant* **13**, 1341–1344 (2020).
55. Varshney, R. V. et al. Can genomics deliver climate-change ready crops? *Curr. Opin. Plant Biol.* **45**, 205–211 (2018).
56. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
57. Wing, R. A., Purugganan, M. D. & Zhang, Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet.* **19**, 505–517 (2018).
58. Varshney, R. Exciting journey of 10 years from genomes to fields and markets: Some success stories of genomics-assisted breeding in chickpea, pigeonpea and groundnut. *Plant Sci.* **242**, 98–107 (2016).
59. Sansaloni, C. et al. Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* **11**, 4572 (2020).
60. Milner, S. G. et al. Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* **51**, 319–326 (2019).
61. Horton, M. et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
62. Ferrero-Serrano, A. & Assmann, S. M. Phenotypic and genome-wide association with the local environment of *Arabidopsis*. *Nat. Ecol. Evol.* **3**, 274–285 (2019).
63. Lasky, J. R. et al. Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Mol. Ecol.* **22**, 5512–5529 (2012).
64. Gutaker, R. et al. Genomic history and ecology of the geographic spread of rice. *Nat. Plants* **6**, 492–502 (2020).
65. Bilinski, P. et al. Parallel altitudinal clines reveal trends in adaptive evolution of genome size in *Zea mays*. *PLoS Genet.* **14**, e1007162 (2018).
66. Lasky, J. R. et al. Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* **1**, e1400218 (2015).
67. Rhoné, B. et al. Pearl millet genomic vulnerability to climate change in West Africa highlights the need for regional collaboration. *Nat. Commun.* **11**, 5274 (2020).
68. Abrouk, M. et al. Fonio millet genome unlocks African orphan crop diversity for agriculture in a changing climate. *Nat. Commun.* **11**, 4488 (2020).
69. Bayer, P. et al. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
70. Danilevicz, M. et al. Plant pangenomics: approaches, applications and advancements. *Curr. Opin. Plant Biol.* **54**, 18–25 (2020).
71. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
72. Brohammer, A. B., Kono, T. J. and Hirsch, C. N. Chapter 2: The maize pan-genome. in *The Maize Genome* (eds Bennetzen, J. et al) (Springer, 2018).
73. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
74. Jayakodi, M. et al. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).
75. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 1–15 (2020).
76. Gao, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).
77. Liu, H. J. & Yan, J. Crop genome-wide association study: a harvest of biological relevance. *Plant J.* **97**, 8–18 (2019).
78. Yang, J., Zhu, J. & Williams, R. W. Mapping the genetic architecture of complex traits in experimental populations. *Bioinformatics* **23**, 1527–1536 (2007).
79. Fan, C. et al. *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor. Appl. Genet.* **112**, 1164–1171 (2006).
80. Zhang, X. et al. Rare allele of *OsPPKL1* associated with grain length causes extra-large grain and a significant yield increase in rice. *Proc. Natl Acad. Sci. USA* **109**, 21534–21539 (2012).
81. Xue, W. et al. Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat. Genet.* **40**, 761–767 (2008).
82. Akakpo, R. et al. The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol.* **226**, 44–49 (2020).
83. Liu, X. et al. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **12**, e1005767 (2016).
84. Huang, M. et al. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience* **8**, giy154 (2019).
85. Voichek, Y. & Weigel, D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat. Genet.* **52**, 534–540 (2020).
86. Kover, P. X. et al. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* **5**, e1000551 (2009).
87. Zaw, H. et al. Exploring genetic architecture of grain yield and quality traits in a 16-way *indica* by *japonica* rice MAGIC global population. *Sci. Rep.* **9**, 19605 (2019).
88. McMullen, M. D. et al. Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
89. Abe, A. et al. Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* **30**, 174–178 (2012).
90. Hammer, G. et al. Models for navigating biological complexity in breeding improved crop plants. *Trends Plant Sci.* **11**, 587–593 (2006).
91. Civelek, M. & Lusis, A. Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* **15**, 34–48 (2014).
92. Rich-Griffin, C. et al. Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. *Trends Plant Sci.* **25**, 186–197 (2020).
93. Libault, M. et al. Plant systems biology at the single-cell level. *Trends Plant Sci.* **22**, 949–960 (2017).
94. Schneider, D. J. & Collmer, A. Studying plant-pathogen interactions in the genomics era: beyond molecular Koch's postulates to systems biology. *Annu. Rev. Phytopathol.* **48**, 457–479 (2010).
95. Whiteman, N. K. & Jander, G. Genome-enabled research on the ecology of plant-insect interactions. *Plant Physiol.* **154**, 475–478 (2010).
96. Turner, T., James, E. K. & Poole, P. S. The plant microbiome. *Genome Biol.* **14**, 209 (2013).
97. Joly-Lopez, Z. et al. An inferred fitness consequence map of the rice genome. *Nat. Plants* **6**, 119–130 (2020).
98. Luo, C. A. R., Fernie & Yan, J. Single-cell genomics and epigenomics: technologies and applications in plants. *Trends Plant Sci.* **25**, 1030–1040 (2020).
99. Efroni, I. et al. Quantification of cell identity from single-cell gene expression profiles. *Genome Biol.* **16**, 9 (2015).
100. Rich-Griffin, C. et al. Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. *Trends Plant Sci.* **25**, 186–197 (2020).
101. Sotelo-Silveira et al. Entering the next dimension: plant genomes in 3D. *Trends Plant Sci.* **23**, 598–612 (2018).
102. Plessis, A. et al. Multiple abiotic stimuli are integrated in the regulation of rice gene expression under field conditions. *eLife* **4**, e08411 (2015).
103. Groen, S. C. et al. The strength and pattern of natural selection on rice gene expression. *Nature* **578**, 572–576 (2020).
104. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
105. Dada, J. & Mendes, P. Multi-scale modelling and simulation in systems biology. *Integr. Biol.* **3**, 86–96 (2011).
106. Xu, K. et al. *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442**, 705–708 (2006).
107. Spindel, J. et al. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116**, 395–408 (2016).
108. Hamblin, M. T., Buckler, E. S. & Jannink, J.-L. Population genetics of genomics-based crop improvement methods. *Trends Genet.* **27**, 98–106 (2011).
109. Spindel, J. et al. Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* **11**, e1004982 (2015).
110. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
111. Mulder, H. A. Is GXE a burden or a blessing? Opportunities for genomic selection and big data. *J. Anim. Breed. Genet.* **134**, 435–436 (2017).
112. Bailey-Serres, J. et al. Genetic strategies for improving crop yields. *Nature* **575**, 109–118 (2019).

113. Kono, T. J. Y. et al. The role of deleterious substitutions in crop genomes. *Mol. Biol. Evol.* **33**, 1669–1678 (2016).
114. Yang, J. et al. Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLoS Genet.* **13**, e1007019 (2017).
115. Liu, Q. et al. Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* **34**, 908–924 (2017).
116. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
117. Wallace, J. G., Rodgers-Melnick, E. & Buckler, E. S. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Ann. Rev. Genet.* **52**, 421–444 (2018).
118. Dwivedi, S. et al. Landrace germplasm for improving yield and abiotic stress adaptation. *Trends Plant Sci.* **21**, 31–42 (2016).
119. Carroll, D. Genome engineering with targetable nucleases. *Ann. Rev. Biochem.* **83**, 409–439 (2014).
120. Urnov, F. et al. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636–646 (2010).
121. Zhang, Y. et al. Transcription activator-like effector nucleases enable efficient plant genome engineering. *Plant Physiol.* **161**, 20–27 (2013).
122. Hua, K. et al. Perspectives on the application of genome-editing technologies in crop breeding. *Mol. Plant* **12**, 1047–1059 (2019).
123. Oliva, R. et al. Broad-spectrum resistance to bacterial blight in rice using genome editing. *Nat. Biotech.* **37**, 1344–1350 (2019).
124. Kwon, C.-T. et al. Rapid customization of Solanaceae fruit crops for urban agriculture. *Nat. Biotech.* **38**, 182–188 (2020).
125. Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).



## Acknowledgements

We thank J. Young Choi, R. Gutaker and A. Kurbidaeva for helpful discussions, and R. Rahni for graphical support. The work is funded by grants from the US National Science Foundation Plant Genome Research Program IOS (grant no. 15-46218), the Zegar Family Foundation (grant no. A168) and the New York University Abu Dhabi Research Institute (grant no. 1205H) (M.D.P.).


## Author contributions

M.D.P. conceived the paper, and wrote it with S.A.J.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00866-3.

**Correspondence** should be addressed to M.D.P.

**Peer review information** *Nature Genetics* thanks Julia Bailey-Serres, Nils Stein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Michael Purugganan

Last updated by author(s): Mar 5, 2021

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☒ | ☐ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☒ | ☐ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | not applicable |
| Data analysis | not applicable |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

not applicable

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | not applicable |
| Data exclusions | not applicable |
| Replication | not applicable |
| Randomization | not applicable |
| Blinding | not applicable |

# Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | not applicable |
| Research sample | not applicable |
| Sampling strategy | not applicable |
| Data collection | not applicable |
| Timing | not applicable |
| Data exclusions | not applicable |
| Non-participation | not applicable |
| Randomization | not applicable |

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | not applicable |
| Research sample | not applicable |
| Sampling strategy | not applicable |
| Data collection | not applicable |
| Timing and spatial scale | not applicable |
| Data exclusions | not applicable |
| Reproducibility | not applicable |
| Randomization | not applicable |
| Blinding | not applicable |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |